https://doi.org/10.24303/lakdoi.2022.30.4.169

Neural Network Language Models as Psycholinguistic Subjects: Focusing on Reflexive Dependency*

Wonil Chung & Myung-Kwan Park**
(Dongguk University)

Chung, Wonil & Park, Myung-Kwan. (2022). Neural network language models as psycholinguistic subjects: Focusing on reflexive dependency. *The Linguistic Association of Korea Journal, 30*(4), 169-190. The purpose of this study is to investigate the reflexive-antecedent dependency resolution accompanying the *wh*-filler-gap dependency resolution in neural network language models (LMs)' sentence processing, comparing the processing result of LMs to the one of humans. To do so, we adopt the psycholinguistic methodology that Fraizer et al. (2015) used for humans. The neural-network language models employed in this study are four LMs: the Long Short-Term Memory (LSTM) trained on large datasets, the Generative Pre-trained Transformer-2 (GPT-2) trained on large datasets, an LSTM trained on small datasets (L2 datasets), and the GPT-2 trained on small datasets (L2 datasets). We found that only the LMs trained on large datasets were sensitive to the dependency between a reflexive and its antecedent matching in gender, but all of the four neural LMs failed to learn reflexive-antecedent dependency accompanying *wh*-filler-gap dependency. Furthermore, we also found that the neural LMs have a learning bias in gender mismatch.

Key Words: reflexive dependency, filler-gap dependency, gender mismatch effect, neural network language model, surprisal

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5B5A16082578).

** The first author is Wonil Chung and the corresponding author, Myung-Kwan Park.

1. Introduction

In this study, we investigate whether the behaviour of neural network language models show incremental syntactic representations reflecting the interaction between the processing of *wh*-filler-gap dependencies and reflexive-antecedent dependencies. To examine how human-like language models (LMs) process both dependencies, we also compare the processing results of LMs with the processing results of native English speakers and Korean English learners.

In on-line sentence comprehension, the human parser establishes dependencies such as *wh*-dependency or reflexive-antecedent dependency between elements encountered in the input string of words. *Wh*-dependency is the dependency between a *wh*-phrase such as *who* or *which* and an empty syntactic position such as subject, direct object, or indirect object, where it is interpreted. Reflexive-antecedent dependency is the dependency between the antecedent noun phrase (NP) and a reflexive pronoun such as *himself* or *herself,* which is typically the later-occurring element.

In sentence processing, *wh*-dependencies (hereinafter referred to as "WhD") and reflexive-antecedent dependencies (hereinafter referred to as "RD") differ from one another. In a WhD, the *wh*-phrase located at the left edge of a clause like (1a) can provide a cue for the existence of an empty direct object position later on. In a RD, reflexive antecedent search is different, because in a sentence like (1b), the reflexive *herself*, which occurs after the antecedent *Lisa*, is overtly marked with morpheme *self*, and there is no indication that *Lisa* is the antecedent until an upcoming reflexive is actually encountered.

(1) a. What did Lisa see __?
b. Lisa saw herself.

Many psycholinguistic studies have revealed that both of these dependency resolution processes occur very rapidly in online reading. Frazier, Ackerman, Baumann, Potter, and Yoshida (2015) showed the antecedent search process was sensitive to syntactic structure, i.e., the presence and location of WhD, when the presence of WhD affects subsequent RD resolution, as in (2), where two NPs such as *which actress* and *Lisa* might be possible antecedents for the reflexive.

(2) Which actress did Lisa imagine to have motivated herself?

In the present study, we examine the processing of constructions like (2), where although *Lisa* is linearly closer to the reflexive *herself*, the only grammatically accessible antecedent for the reflexive *herself* is the more distant *wh*-NP, *which actress,* comparing the neural network language models' language processing with human language processing. That is, the goal of this paper is to examine how well the neural network language models perform the interaction between the processing of *wh*-dependencies and reflexive-antecedent dependencies, compared to human language processing.

The neural network language model is a language model that mechanically implements human language processing using computational natural language processing (NLP) technology, and can be defined as a probability distribution for a word sequence. Neural network language models that use neural sequence models of various kinds to derive sentence representations have been able to achieve impressive results on some tasks, using experimental techniques developed in the field of psycho/neurolinguistics to study language processing in the human mind. (Elman, 1990; Sutskever et al., 2014; Goldberg, 2017; Peters et al., 2018; Devlin et al., 2018; Goodkind & Bicknell, 2018; Wilcox et al., 2018; Aurnhammer & Frank, 2019; Hu et al., 2020; Hao et al., 2020; Wilcox et al., 2020; Da Costa & Chaves, 2020; Chaves & Richter 2021; Ryu & Lewis 2021; Wilcox et al., 2021).

This approach using experimental techniques was introduced by Linzen, Dupoux, and Goldberg (2016) using the agreement prediction task (Bock & Miller 1991) to study the hierarchical morphosyntactic dependency of recurrent neural networks (RNNs). Subsequently, Gulordava et al. (2018) revealed that subject-verb agreement dependency is learnable from language modeling objective. This approach has extended to other grammatical phenomena such as filler-gap dependencies showing positive results (Chowdhury & Zamparelli, 2018; Wilcox et al., 2018), and reflexive dependencies showing negative results (Marvin & Linzen 2018).

In this study, it is focused on whether the neural network language models show evidence for incremental syntactic representation reflecting the interaction between the processing of *wh*-filler-gap dependencies and reflexive-antecedent dependencies, considering the processing results of the neural network language model on *wh*-filler-gap dependency or reflexive dependency in previous studies. In order to conduct the experiment, we consider the neural network language model as a subject of a psycho/neurolinguistic experiment. Furthermore, to compare neural network language models as models of human sentence processing, we compute the surprisal or log inverse probability the language models assign to stimuli used in the self-paced reading or eye-tracking experiments. In psycho/neurolinguistics, reaction time per word, as a measure of the word-by-word difficulty of sentence processing, is taken to reflect the extent to which humans expect a word in context. The surprisal value of Surprisal Theory is known to correlate with human processing difficulty and provides a link between psycho/neurolinguistic modeling and neural network language modeling (Hale, 2001; Levy, 2008).

As the experimental method of this study, first, we collect surprisal values estimated by the LSTM (Long Short-Term Memory) (Gulordava et al., 2018) language model and the GPT-2 (Generative Pre-trained Transformer 2) (Radford et al., 2019), which are pre-trained autoregressively on a large amount of data. Second, to investigate the interaction between the processing of *wh*-filler-gap dependency and reflexive- antecedent dependency during sentence processing, we collect the reaction times (RT) measured by the self-paced reading (SPR) experiments of Korean English learners using stimuli used in the eye-tracking (ET) experiments of native English speakers in Frazier et al. (2015). Third, we compare the surprisal estimated from LSTM and GPT-2 at critical region, reflexive such as *himself* or *herself*, with RTs from human.

This research paper is organized as follows. Chapter 2 describes previous studies on linguistic theory and neural network language models, Chapter 3 describes experimental analyses of WhD and RD processing in human or neural network language models, and Chapter 4 describes discussion and conclusion of aspects of neural network language models in language processing.

2. Previous Studies

2.1. Theoretical Background of Linguistics

In the present study, it is investigated whether the resolution of a WhD establishes a new candidate antecedent in the searched representation during the resolution of a RD in sentence processing, and examine the interaction between WhD and RD. In psycho/neurolinguistics, sentence processing studies indicate that resolving a WhD is an active process. Upon encountering a *wh*-phrase, the parser activates the dependency processing, detecting an incoming position at which resolving a WhD would be grammatically accessible (Stowe, 1986; Traxler & Pickering, 1996; Phillips, 2006). It is also known that upon encountering the reflexive, the parser attempts to link the reflexive to grammatically accessible antecedents in the early stages of sentence processing (Nicol & Swinney, 1989; Sturt, 2003; Jäger et al., 2015).

When considering reflexive dependencies in (3), the reflexive *himself* co-refers with its nearest potential antecedent *the man,* not with *Anne*. However, in (4), *wh*-phrase *which man* is understood as the subject of the non-finite embedded clause *to have motivated himself*, as in (3), but it is distant from the embedded clause subject position after *expect*. In the context containing such non-finite embedded clause, the antecedent of *himself* becomes *wh*-phrase *which man* instead of the linearly closer NP *Anne*. If *Anne* was chosen as the reflexive antecedent in (3) or (4), sentences (3) and (4) would be unacceptable due to the gender mismatch between the male reflexive *himself* and the female name *Anne.*

(3) Anne$_i$ expected the man$_j$ to have motivated himself$_{*i/j}$.
(4) Which man$_i$ did Anne$_j$ expect to have motivated himself$_{i/*j}$?

The result of the *wh*-dependency resolution influences reflexive dependency resolution. In (4), without WhD resolution, the closest potential candidate antecedent *Anne* mismatches with the reflexive *himself* in gender, leading to processing difficulty and slow reading time in sentence processing (Sturt, 2003). However, if WhD is activated, the new candidate antecedent for the reflexive *himself* will be established. In this case, the antecedent *which man* that is closer than the ungrammatical antecedent *Anne* will be associated with the reflexive *himself.* Therefore, the parser will experience a gender mismatch effect when the reflexive mismatch in gender with an ungrammatical but linearly close candidate antecedent like *Anne*. In Frazier et al. (2015)'s eye-tracking experiments, humans were sensitive to gender mismatch, detecting the WhD resolution during the RD resolution. That is, the parser's reflexive antecedent search was sensitive to syntactic structure such as the presence and location of a WhD.

In online sentence comprehension using materials with syntactic structure like (5), Sturt (2003) reported the reflexive was read more slowly when the gender of the reflexive mismatches the grammatically accessible antecedent, the stereotypical gender, (e.g. *herself vs surgeon*) than when it matched (e.g. *himself vs surgeon*) in an eye-tracking study. Dillon et al. (2013) found effects of the accessible antecedent without any effect of the inaccessible antecedent using materials of Sturt (2003) in an eye-tracking study. Xiang et al. (2009) found P600 at the reflexive that does not match the stereotypical gender of the grammatically accessible antecedent in an ERP study.

(5) The surgeon who treated Jonathan had pricked himself.

Contrary to the claim that reflexive antecedent search is structurally sensitive, in the computational model based on the retrieval cues of an antecedent search (Lewis & Vasishth, 2005), because the parser is able to consider all possible candidate antecedents for the reflexive while interacting with non-structural cues, the mismatch-mismatch conditions lead to slowdown in reading time experiencing difficulty in the absence of a gender-matching candidate antecedent. Jäger et al. (2015) and Jäger et al. (2020) found both candidate antecedents affect reading times at reflexive in cue-based retrieval model that is not constrained by syntactic structure.

2.2. Neural Network Language Models

The neural network language model is a model that assigns probability to word sequences and predicts the next word using previous words in context. If this presents as a conditional probability, the predicted value of *himself* in the sentence *John liked himself* can be presented as $P(himself \mid John,\ liked)$. Based on this probability distribution, many of the previous studies suggest that the time it takes humans to read a word can be predicted by estimating the word's probability in context, that is, real-time language comprehension involves predictions about upcoming words in context. In general, psychometric predictive power by using surprisal value or log inverse probability from a neural network language model turns out to be correlate with online processing measures including self-paced reading times, gaze duration in the eye-tracking studies, and N400 measures in EEG studies (Smith & Levy, 2013; Frank et al., 2015).

In order to assess the language learning/processing performance of neural network language models, recent studies have followed controlled psycholinguistic-style testing for grammatical knowledge (Marvin & Linzen, 2018; Futrell et al., 2018; Van Schijndel & Linzen, 2018; Wilcox et al., 2020; Linzen & Baroni, 2021). Furthermore, recent studies have evaluated neural network language models by assessing the predictive power of the surprisal that each model assigns to stimuli used in experiments of humans reading (Goodkind & Bicknell, 2018; Wilcox et al., 2018; Aurnhammer & Frank, 2019; Da Costa & Chaves, 2020; Hu et al., 2020; Hao et al., 2020; Chaves, 2020; Ryu & Lewis, 2021; Chaves & Richter 2021; Wilcox et al., 2021).

The surprisal or negative log-conditional probability known to predict human incremental processing difficulty is estimated by the probability value of occurrence of a word (w) within a given preceding context (c). The following formula indicates the surprisal, $S(w_i)$ of a sentence's i-th word w_i.

$$S(w_i) = -\log p(w_i|c) = -\log(p(w_i|w_1 \ldots w_{i-1}))$$

In Surprisal Theory (Hale, 2001; Levy, 2008), the surprise of a word is the degree of expectation that is linearly related to the difficulty of the word, so a word with a high surprisal has a lower expectation in context than a word with a low surprisal. Many language model studies using experimentally controlled sentences and the surprisal value have investigated whether neural network language models are able to learn and generalize about syntactic knowledge. Hu et al. (2020) investigated whether neural language models learn and generalize human-like syntactic knowledge on 6 syntactic circuits[1] including 34 English-language test suites covering a wide range of syntactic phenomena, testing 5 model types (LSTM, ON-LSTM, RNNG, GPT-2, and n-gram) and 4 types of data sizes (1M, 5M, 14M and 42M tokens). They found significant differences in syntactic generalization scores by model architecture, and also a greater effect of model inductive bias than training data size on syntactic generalization score. Model inductive biases have little effect on performance on Licensing including Negative Polarity Item Licensing (NPI) and Reflexive Pronoun Licensing, both from Marvin and Linzen (2018). Within syntactic phenomena, there was little effect of dataset size on syntactic generalization score except for Agreement. Pre-trained GPT-2 outperform all other models on each syntactic phenomenon including Licensing. In their GPT-2 results, the influence of model architecture relative to data size offers another striking example. While GPT-2 trained on 14M tokens and GPT-2 trained on 42M tokens achieve almost the same

1) Agreement includes Subject-Verb Number Agreement. 2. Licensing includes Negative Polarity Item and Reflexive Pronoun. 3. Garden-Path Effects include Main Verb/Reduced Relative Clause and NP/Z Garden-paths. 4. Gross Syntactic Expectation includes Subordination. 5. Center Embedding. 6. Long-Distance Dependencies include Filler-gap Dependencies and Cleft.

syntactic generalization score as the pre-trained GPT-2 trained on 40GB of web text (Radford et al. 2019), GPT-2 trained on smaller dataset (1M or 5M tokens) showed the poor performance that may be due to overparameterization.

Wilcox et al. (2018) studied to investigate whether LSTM language model represents filler-gap dependencies, using experimentally controlled sentences and estimating the surprisal value from the language model. They found LSTM language models learned and generalized about empty syntactic positions, using two models, Google model trained on 0.8B words (Jozefowiez et al., 2016) and Gulordava model trained on 90M words (Gulordava et al., 2018).

Futrell et al. (2018) studied to investigate whether how well LSTM language model learns and represents incremental syntactic state and grammatical dependency, employing the methods of controlled psycholinguistic experiment. They found although LSTM language model represented and maintained incremental syntactic state, language models did not generalize in the same way as humans. Furthermore, their language model did not learn the appropriate grammatical dependency such as reflexive pronouns mismatching the antecedent's stereotypical gender or negative polarity items. In reflexive pronoun binding, one of two LSTMs, GRNN (Generalized Regression Neural Network) (Gulordava et al., 2018) trained on 90M tokens of English Wikipedia, did not show a reliable effect of stereotypical gender. The other of two LSTMs, JRNN trained on 0.8B words, had higher surprisal at reflexive pronouns mismatching the stereotypical gender antecedent than at pronouns matching the stereotypical gender antecedent. In particular, although humans do not consider antecedents outside the binding domain as antecedents for reflexives (Sturt, 2003; Xiang et al., 2009; Dillon et al., 2013), LSTM language model, JRNN, was influenced by intervener gender due to lower surprisal when the intervener matches reflexive gender among conditions where true antecedent gender mismatches reflexive gender in the sentence *The lumberjack who is related to the hairdresser cut herself*. GRNN did not show a reliable effect of stereotypical gender.

In the present study, we examine whether the neural network language models such as LSTM (Gulordava et al., 2018) and the GPT-2 (Radford et al., 2019) which are autoregressive pre-trained language models learn and represent the interaction between *wh*-filler-gap dependencies and reflexive-antecedent dependencies in sentence processing.

3. Experiments

In this study, in order to examine whether the neural network language model(LM) can process the reflexive-antecedent dependency like native speakers or human, selecting the filler-gap dependency as the antecedent of a reflexive dependency, psycholinguistic methods have been emplyed and experimental materials from Fraizer et al. (2015) have been adopted. For LM processing data to compare data type and data size, we collected surprisal values at critical region (e.g., *herself* or *himself*) for four pre-trained language models: an LSTM (Gulordava et al., 2018) trained on large datasets (90M words), the GPT-2 (Radford et al., 2019) trained on large datasets (800M words), an LSTM trained on the small datasets (L2 datasets), and the GPT-2 trained on the small datasets (L2 datasets). L2 datasets (7900K words) consist of English textbooks which Korean learners of English can potentially encounter in their English learning. For L2ers processing data to compare with native English speakers, we use reaction times (RT) at critical region. We collected them from late learners with high proficiency in L2 English (scores on TOEIC Test: 850-985) in self-paced reading paradigm. For native speakers processing data, we used the results of Fraizer et al.'s study. In order to investigate how well LMs or L2ers process the interaction between processing of reflexive-antecedent dependency and the filler-gap dependency like native speakers', we performed two-way ANOVA with four conditions as two within-items factors (wh-phrase & local NP) for statistical analyses.

3.1. Materials

In this study, four experiments were adopted from Fraizer et al. (2015), which conducted examples as shown in Table 1. In each experiment, materials were constructed in a two-by-two factorial design with *wh*-NP factor and local NP factor, consisting of 4 conditions.

Each sentence consists of a matrix clause involving a *wh*-NP at the left edge of a complex sentence and an embedded clause containing a reflexive pronoun, and has the gender match/mismatch of the reflexive pronoun with the *wh*-NP and the linearly closer matrix-clause subject. Each condition consisted of 24 sentences. The critical region was the reflexive pronoun (e.g., *herself* or *himself*). Experiments differ in whether the embedded clause was non-finite (1 and 3) or finite (2 and 4), and in whether the target *wh*-NP was the subject of embedded clauses (1 and 2) intervened between the reflexive and its closest overt antecedent (1 and 2) or the *wh*-NP was the subject of matrix clauses (3 and 4).

Table 1. An Example of Experimental Materials

EXP 1: Non-finite embedded clauses
• Wh-NP-match; local NP-match/mismatch Which actress did Lisa/James imagine to have motivated herself • Wh-NP-mismatch/ local NP-match/mismatch Which actress did James/Lisa imagine to have motivated himself
EXP 2; Finite embedded clauses
• Wh-NP-match; local NP-match/mismatch Which actress did Lisa/James imagine had motivated herself • Wh-NP-mismatch/ local NP-match/mismatch Which actress did James/Lisa imagine had motivated himself
EXP 3: Non-finite embedded clauses
• Wh-NP-match; local NP-match/mismatch Which actress imagined Lisa/James to have motivated herself • Wh-NP-mismatch/ local NP-match/mismatch Which actress imagined James/Lisa to have motivated himself
EXP 4: Finite embedded clauses
• Wh-NP-match; local NP-match/mismatch Which actress imagined Lisa/James had motivated herself • Wh-NP-mismatch/ local NP-match/mismatch Which actress imagined James/Lisa had motivated himself

3.2. Neural Network Language Models' Sentence Processing

3.2.1. GPT-2 Models

The results of surprisal value of each condition in Figure 1 show the difference between conditions in each experiment, and also the difference between the L1_GPT-2 and the L2_GPT-2 models. In L1_GPT-2, the conditions where the *wh*-NP and reflexive match (e.g. *which actress* and *herself*) were lower in surprisal than the conditions where the *wh*-NP and reflexive were mismatched (e.g. *which actress* and *himself*) in Experiment 1 and 2. In Experiment 3 and 4, the conditions where the local-NP and reflexive match (e.g. *Lisa* and *herself* or *James* and *himself*) were lower in surprisal than the conditions where the local NP and reflexive were mismatched (e.g. *Lisa* and *himself*). Globally, *wh*-NP match conditions were lower in surprisal than *wh*-NP mismatch conditions.

However, in L2_GPT-2, the conditions where the *wh*-NP and reflexive match (e.g.

which actress and *herself*) were slightly lower in surprisal than the conditions where the *wh*-NP and reflexive were mismatched (e.g. *which actress* and *himself*) in 4 Experiments.

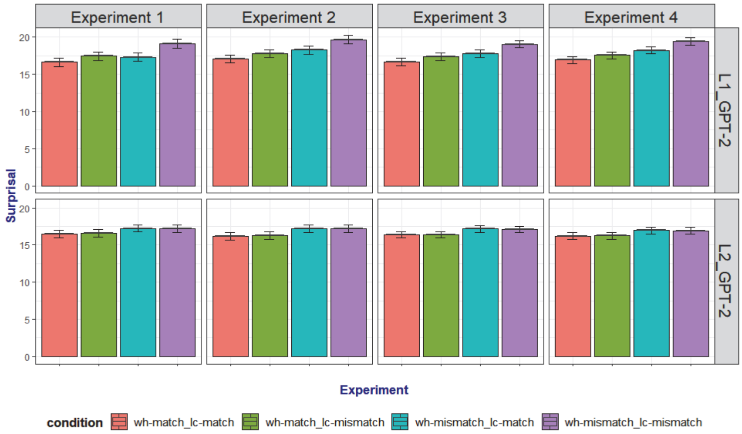

Figure 1. Surprisal from GPT-2 at Reflexive in Each Condition

For statistical analysis, all surprisals were submitted to 2 × 2 Analyses of Variance, aggregating by item. The results of the effect of each factor (i.e. *wh*-NP or local NP) and interaction between *wh*-NP and local NP are listed in Table 2. In L1_GPT-2, wh-NP factor was significant in all experiments due to sensitive to *wh*-NP, common noun (e.g. *actress*), during the reflexive antecedent search process. Namely, at reflexive region the conditions in which the gender of the reflexive pronoun mismatched with that of the *wh*-NP were higher surprisals than the conditions in which the genders matched. Similarly, local NP effect was significant in the embedded clause which was non-finite in Experiment 1 and 3, due to sensitive to local-NP, proper name (e.g. *Lisa* or *James*). In Experiment 2 and 4, local NP effect was marginally significant in the embedded clause which was finite. In contrast to L1_GPT-2, there was no significant effect in any experiment in L2_GPT-2.

Table 2. ANOVA results for 4 Experiments in GPT-2

	Factor	EXP 1	EXP 2	EXP 3	EXP 4
L1_GPT-2	wh-NP	4.48*	8.00**	7.89**	10.4**
	local NP	5.32*	3.65†	4.06*	3.25†
	wh*local	-	-	-	-

	Factor	EXP 1	EXP 2	EXP 3	EXP 4
L2_GPT-2	wh-NP	-	3.36†	3.35†	-
	local NP	-	-	-	-
	wh*local	-	-	-	-

3.2.2. LSTM Model

Mean surprisal values of each condition in Figure 2 show the difference between conditions in each experiment, and also the difference between the L1_LSTM and the L2_LSTM models. In L1_LSTM, the conditions where the *wh*-NP and reflexive match (e.g. *which actress* and *herself*) were slightly lower in surprisal than the conditions where the *wh*-NP and reflexive were mismatched (e.g. *which actress* and *himself*) in Experiment 1 and 2. In Experiment 3 and 4, the conditions where the local-NP and reflexive match (e.g. *Lisa* and *herself* or *James* and *himself*) were lower in surprisal than the conditions where the local-NP and reflexive were mismatched (e.g. *Lisa* and *himself*). In contrast to L1_GPT-2, in L1_LSTM, local-NP match conditions were lower in surprisal than local-NP mismatch conditions.

However, in L2_LSTM, the conditions where the *wh*-NP and reflexive match (e.g. *which actress* and *herself*) were higher in surprisal than the conditions where the *wh*-NP and reflexive were mismatched (e.g. *which actress* and *himself*) in 4 Experiments. We assumed that L2_LSTM did not learn binding, which characterizes the syntactic restrictions on reflexive and their antecedents (common noun or proper name), so we do not examine it further in statistical analysis.

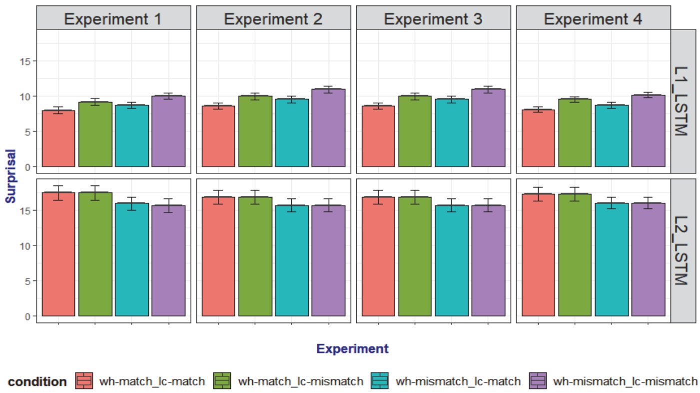

Figure 2. Surprisal from LSTM at reflexive in each condition

Table 3. ANOVA Results for 4 Experiments in LSTM

	Factor	EXP 1	EXP 2	EXP 3	EXP 4
L1_LSTM	wh-NP	-	4.14*		4.09*
	local NP	7.57**	9.19**	13.5***	17.2***
	wh*local	-	-	-	-
L2_LSTM	wh-NP	-	-	-	-
	local NP	-	-	-	-
	wh*local	-	-	-	-

For statistical analysis, all surprisals were submitted to 2 × 2 Analyses of Variance, aggregating by item. The results of the effect of each factor (i.e. *wh*-NP or local NP) and interaction between *wh*-NP and local NP are listed in Table 3. In L1_LSTM, local NP effect was significant in all experiments due to sensitive to local-NP, proper name (e.g. *Lisa* or *James*), during the reflexive antecedent search process. In other words, at reflexive region the conditions in which the gender of the reflexive pronoun mismatched with that of the local-NP were higher surprisals than the conditions in which the genders matched. However, *wh*-NP effect was significant in the embedded clause which was finite in Experiment 2 and 4, due to sensitive to wh-NP, common noun (e.g. *actress*).

3.3. Korean English Learners

Korean English learners participated in these experiments (34 in EXP 1, 31 in EXP 2, 20 in EXP 3, and 22 in EXP 4), and they were undergraduates (mean age 24.4 in EXP 1, 24.8 in EXP 2, 24.8 in EXP 3, and 24.7 in EXP 4).

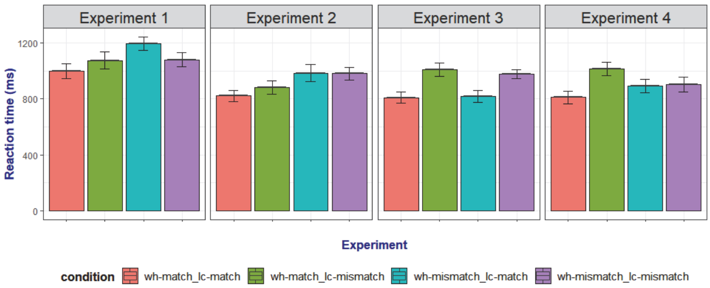

Figure 3. RT from Korean English Learners at Reflexive in Each Condition

The results of RT of each condition in Figure 3 show the difference between conditions in each experiment. The conditions where the *wh*-NP and reflexive were mismatched (e.g. *which actress* and *himself*) were slower RT than the conditions where the *wh*-NP and reflexive were matched (e.g. *which actress* and *herself*) in Experiment 1 and 2. In contrast, in Experiment 3 and 4, the conditions where the local-NP and reflexive were mismatched (e.g. *Lisa* and *himself* or *James* and *herself*) were slower RT than the conditions where the local-NP and reflexive were matched (e.g. *Lisa* and *herself* or *James* and *himself*).

For statistical analysis, all RTs were submitted to 2 × 2 Analyses of Variance, aggregating by item. The results of the effect of each factor (i.e. *wh*-NP or local NP) and interaction between *wh*-NP and local NP are listed in Table 4. In Experiment 1, *wh*-NP factor revealed marginal effect (p = 0.067) and interaction effect was marginal (p = 0.077). In Experiment 2, *wh*-NP factor was significant ($p < 0.01$). In contrast to Experiment 1 and 2, in Experiment 3 showed local NP effect was significant ($p < 0.001$), and also in Experiment 4 showed local NP effect was significant ($p < 0.05$). These results showed *wh*-NP served as antecedent of reflexive in Experiment 1 and 2, and local NP acted as antecedent of reflexive in Experiment 3 and 4.

Table 4. ANOVA Results for 4 Experiments in L2ers

factor	EXP 1	EXP 2	EXP 3	EXP 4
wh-NP	3.44†	7.20**		
local NP	-		19.79***	5.01*
wh*local	3.20†			3.89†

3.4. Native Speakers

Figure 4 shows mean RT reported by Fraizer at al. (2015) at critical region, reflexive (e.g. *himself* or *herself*), in each condition in each experiment. In Experiment 1 and 2, the conditions where the *wh*-NP and reflexive were mismatched (e.g. *which actress* and *himself*) revealed slower RT than the conditions where the *wh*-NP and reflexive were matched (e.g. *which actress* and *herself*).

In Experiment 3 and 4, the conditions where the local-NP and reflexive were mismatched (e.g. *Lisa* and *himself* or *James* and *herself*) showed slower RT than the conditions where the local-NP and reflexive were matched (e.g. *Lisa* and *herself* or *James* and *himself*).

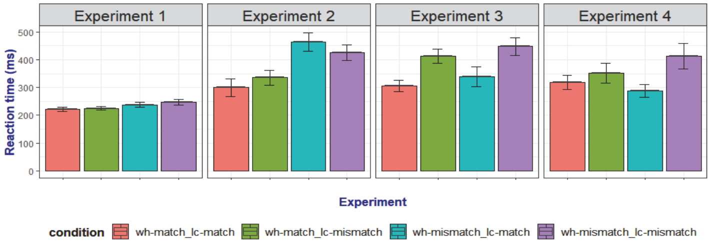

Figure 4. RT from Native Speakers at Reflexive in Each Condition

3.5. Comparison of Effects in Human or LM

In native English speakers, the process of reflexive-antecedent resolution is sensitive to the presence of a *wh*-filler-gap dependency which was the grammatically licit antecedent of the reflexive in Experiment 1 and 2. In Experiment 3 and 4, there was a main effect of local NP at the critical region due to faster RT in the gender matched conditions than gender mismatched condition. Likewise, in Korean English learners, reflexive-antecedent resolution in Experiment 1 and 2 was sensitive to *wh*-NP which is the grammatically licit antecedent of the reflexive, whereas reflexive-antecedent resolution in Experiment 3 and 4 was sensitive to the local NP serving as the sole grammatically licit antecedent.

Table 5. Comparison of Effects in Human or LM

		Human		GPT-2		LSTM	
EXP	factor effect	L1	L2	L1	L2	L1	L2
EXP 1	wh-NP	✔	†	✔	-	-	-
	local NP	-	-	✔	-	✔	-
	wh-phrase*local NP	-	†	-	-	-	-
EXP 2	wh-phrase	✔	✔	✔	†	✔	-
	local NP	-	-	†	-	✔	-
	wh-phrase*local NP	-	-	-	-	-	-
EXP 3	wh-phrase	-	-	✔	†	-	-
	local NP	✔	✔	✔	-	✔	-
	wh-phrase*local NP	-	-	-	-	-	-
EXP 4	wh-phrase	-	-	✔	-	✔	-
	local NP	✔	✔	†	-	✔	-
	wh-phrase*local NP	-	†	-	-	-	-

✔: significant effect; †: marginal effect

In contrast to humans' response to the interaction of two non-local dependencies, *wh*-filler-gap dependencies and reflexive-antecedent dependencies, neural network language models revealed different results. The L2_GPT-2 and L2_LSTM, trained by English textbooks published in Korea, showed no effect in any experiment. Unlike L2_GPT-2 and L2_LSTM, the results of two models, the results of L1_GPT-2 and L1_LSTM showed a unique contrast. Regardless of the grammatically licit antecedent for reflexive, while L1_GPT-2 was sensitive to *wh*-NP which consists of *which* and common noun, L1_LSTM was sensitive to local NP which consists of proper name, regardless of the gender match/mismatch of antecedent for reflexive. Furthermore, while L1_GPT-2 was sensitive to local NP which consists of a proper name in the embedded clause which was non-finite in Experiment 1 and 3, L1_LSTM was sensitive to *wh*-NP which consists of *wh*-NP which consists of *which* and common noun, in the embedded clause which was finite in Experiment 2 and 4.

4. Discussion and Conclusion

The experiments in this paper have been to investigate whether like native speakers, the neural network language models such as the LSTM LM (Gulordava et al., 2018) and the GPT-2 LM (Radford et al., 2019) can process the reflexive-antecedent dependency at issue that accompanies the wh-filler-gap dependency, by using the controlled experimental materials from Fraizer et al. (2015). For native speakers, the results of the eye-tracking text-reading experiments reported by Fraizer et al. in their Experiments 1 and 2 show that the parser selected the grammatical but linearly distant antecedent (i.e., a licit *wh*-NP) as the reflexive antecedent during the reflexive antecedent search. In their Experiments 3 and 4, while the *wh*-NP did not serve as a grammatically accessible antecedent for the reflexive during the reflexive antecedent search, the local NP served as a grammatically licit antecedent. Meanwhile, as reported in Chung and Park (2018), for L2ers, while their results of the four experiments were shown to be analgous to native speakers', they did not consider antecedents outside the binding domain as antecedents for reflexives.

Unlike the results of L1 and L2 human processing for the reflexive-antecedent dependence accompanying the wh-filler-gap dependency, the neural network language models in the present experiments failed to choose a grammatically licit antecedent for reflexive resolution, failing to select a distant wh-filler as the antecedent of a reflexive dependency. Neither of the two L2_GPT-2 and L2_LSTM LMs trained on the small datasets (L2 datasets) captured an interaction between the processing of both *wh*-filler-gap and reflexive-antecedent dependencies. We suspect that the poor performances in reflexive resolution by the L2_GPT-2 and L2_LSTM LMs trained on small L2 datasets was due to the rare attestations of binding sentences in the dataset. The L2 datasets, which were collected from English textbooks published in Korea, do not have enough wh-NPs as well as proper names and common nouns that can serve as antecedents of the reflexives in the test dataset. Furthermore, as mentioned above, the poor performance of the task at issue may have been due to the over-parameterization of the LMs in the training stages (Hu et al., 2020).

By contrast, the L1_GPT-2 and L1_LSTM LMs trained on large datasets showed different results. First, the L1_GPT-2 LM was sensitive to the presence of *wh*-NPs, regardless of whether they served as a grammatically licit antecedent or a grammatically illicit antecedent. In contrast to the L1_GPT-2 LM, the L1_LSTM LM was sensitive to the presence of local NPs, regardless of whether they served as a grammatically licit or illicit antecedent. To identify the reason for this difference, we performed an additional comparison of gender match/mismatch for reflexives. Both LMs showed that gender mismatch conditions were higher in surprisal than gender match conditions.However,there was a difference between the L1_GPT-2 and the L1_LSTM LMs. While the L1_GPT-2 LM showed a significant effect, $t(45.89)=-2.3509$, $p < 0.05$, in the common noun condition like *The actress had motivated herself/*himself*, the L1_LSTM showed a significant effect, $t(45.20)=-3.8738$, $p < 0.001$, in the proper name condition like *Lisa had motivated herself/*himself*. We suspect that this is due to the difference in the architecture and the size of training datasets between the GPT-2 LM (Radford et al., 2019) trained on large datasets (800M words) and the LSTM LM (Gulordava et al., 2018) trained on large datasets (90M words).

Furthermore, the L1_GPT-2 and the L1_LSTM LMs showed different gender mismatch effects depending on sentence structures. When the embedded clause was non-finite, the L1_GPT-2 LM was sensitive to a local NP which is composed of a proper name (*Lisa* or *James*). By contrast, when the embedded clause was finite, the L1_LSTM LM was sensitive to a *wh*-NP which is composed of a common noun (*actress* or *actor*). We suggest that this difference is also due to the architecture and the amount of training datasets.

Recent neural-network language models are often described as language learners that lack innate biases and induce all their cognitive abilities from given learning data (Fodor

& Pylyshyn, 1988; Pinker & Prince, 1988; Christiansen & Chater, 1999). If so, the successful syntactic performances of such neural network language models may be taken to indicate that their human-like syntactic ability can be acquired through simple statistical learning. However, any learning theory will dictate that the concept of a tabula rasa is inconsistent in practice. Therefore, a useful learner including a neural language model must have certain innate or pre-equipped biases that drive it to favor some possible generalizations over others (Mitchell, 1980). Neural LMs also certainly have biases arising from their initial weights and architectural features, which incorporate assumptions of temporal invariance, attention, encoding and decoding modules, and other architectural elements.

In this study, we have found that like native speakers, human L2ers processed reflexive dependency accompanying *wh*-filler-gap dependency, successfully selecting a grammatically licit antecedent for reflexives, but the neural network language models adopted in this paper did not capture native-like gender mismatch in reflexive resolution. Such neural LMs were influenced by their architecture features and the size of training datasets that resulted in inducing their internal and data biases. In conclusion, we note that neural network language models have reflexive learning biases in light of gender match/mismatch in reflexive-antecedent dependency accompanying wh-filler-gap dependency. In the future, instead of a large-scale pre-trained language model, it is necessary to conduct a follow-up research to confirm the performance of the language model by learning linguistic phenomena centered by linguists.

References

Aurnhammer, C., & Frank, S. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia, 134*, 107-198.

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology, 23*(1), 45-93.

Chaves, R. P. (2020). What don't RNN language models learn about filler-gap dependencies? *Proceedings of the Society for Computation in Linguistics, 3*(1), 20-30.

Chaves, R. P., & Richter, S. N. (2021). Look at that! BERT can be easily distracted from paying attention to morphosyntax. *Proceedings of the Society for Computation in Linguistics, 4*(1), 28-38.

Chowdhury, S. A., & Zamparelli, R. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics,* 133-144.

Christiansen, M., & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive Science, 23*, 417–37.

Chung, W., & Park, M.-K. (2018). Are Korean English learners structure-sensitive in reflexive resolution? *The Asian International Journal of Life Sciences, 15*(4), 2905-2919.

Da Costa, J., & Chaves, R. (2020). Assessing the ability of Transformer-based Neural Models to represent structurally unbounded dependencies. *Proceedings of the Society for Computation in Linguistics, 3*(1), 189-198.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language, 69*(2), 85-103.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Fodor, J., & Pylyshyn, Z. (1988), Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3-71.

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *Journal of Neuroscience, 35*(2), 485-494.

Frazier, M., Ackerman, L., Baumann, P., Potter, D., & Yoshida, M. (2015). Wh-filler-gap dependency formation guides reflexive antecedent search. *Frontiers in Psychology, 6*, 01504.

Futrell, R., Wilcox, E., Morita, T., & Levy, R. (2018). RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329.*

Goldberg, Y. (2017). *Neural network methods for natural language processing*. San Francisco, CA: Morgan & Claypool.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018),* 10-18.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138.*

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics,* 1-8.

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *arXiv preprint arXiv:2009.03954*.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2015). Retrieval interference in reflexive processing: Experimental evidence from Mandarin, and computational modeling. *Frontiers in psychology, 6*(617).

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language, 111*(104063).

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu. Y. (2016). Exploring the limits of language modeling. *arXiv 1602.02410.*

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126-1177.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*(3), 375-419.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics, 7*, 195-212.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics, 4*, 521-535.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031.*

Mitchell, T. M. (1980). The need for biases in learning generalizations. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Rutgers University, New Brunswick, NJ

Nicol, J., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of psycholinguistic research, 18*(1), 5-19.

Peters, M., Neumann, E. M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv:1802.05365.*

Phillips, C. (2006). The real-time status of island phenomena. *Language, 82*, 795–823.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*, 73–193.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Ryu, S. H., & Lewis, R. L. (2021). Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. *arXiv preprint arXiv:2104.12874*.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302-319.

Stowe, L. A. (1986). Parsing WH-constructions: Evidence for on-line gap location. *Language and Cognitive Processes, 1*(3), 227-245.

Sturt, P. (2003). A new look at the syntax-discourse interface: The use of binding principles in sentence processing. *Journal of Psycholinguistic Research, 32*(2), 125-139.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems, 27*, 3104–3112.

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, *35*(3), 454-475.

Van Schijndel, M., & Linzen, T. (2018). A neural model of adaptation in reading. *arXiv preprint arXiv:1808.09930.*

Wilcox, E., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Wilcox, E., Futrell, R., & Levy, R. (2021). Using computational models to test syntactic learnability. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43,* 1-88.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.
Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language, 108*(1), 40-55.

Wonil Chung
Research Professor
Department of English Language and Literature
Dongguk University
30 Phildongro 1-gil, Jung-gu
Seoul 04620, Korea
Email: wonilchung@naver.com

Myung-Kwan Park
Professor
Department of English Language and Literature
Dongguk University
30 Phildongro 1-gil, Jung-gu
Seoul 04620, Korea
Email: parkmk@dgu.edu

Received on November 15, 2022
Revised version received on December 15, 2022
Accepted on December 31, 2022